Learning from Probabilistic Class Labels

Padhraic Smyth
Jet Propulsion Laboratory '238-420
California institute of Technology
4800 oak Grove Drive
Pasadena, CA 91109
[pjs@galway.jpl.nasa.gov]

Abstract

In this paper the problem of supervised learning is addressed where class labels in the training data do not correspond to ground truth, but instead are subjective estimates of class membership provided by a domain expert. This is a practical problem in application such as remote-sensing and medical diagnosis where labelling of the feature data may take place in a subjective manner some time after the original data was measured. In particular the case where the labels can be interpreted as estimates of posterior class probabilities is examined. A variety of labelling strategies exist, of which oracle-based, probabilistic, and maximum a posteriori (MAP) labelling are among the most interesting. Basic relationships between these strategies arc established. For example, MAP labelling does not provide enough information to the learning algorithm to properly recover class probability estimates, however, it does permit the algorithm to learn a minimum-error classifier. The role of side-information is briefly discussed, where the labeller may be using additional side information (not present in the measured features) to label the data.

The practical question of how probabilistic labels might be best used with atypical learning algorithm is addressed. in fact, the modifications to existing learning algorithms are generally straightforward, particularly for loss-function based discriminants (such as multi-layer perceptron models). For parametric models it can be shown that asymptotically consistent estimators exist [], 2]: intuitively the approach is that attaining sample isdivided up between the classes in proportion to its class label weight. Empirical results on test data sets show that probabilistic labelling universally outperforms the more conventional deterministic labelling, in terms of both error rate and posterior probability approximation. While the improvement in error rate is typically slight, the improvement in probability approximation capabilities can be very substantial (orders of magnitude in mean squared error). In particular, sigmoid-based network models with a mean squared error loss function appear to take the greatest advantage of the probabilistic labelling - this can be explained by relating the network model to logistic discrimination for general exponential families [3]. in conclusion it is noted that, in the real-world, elicitation of accurate and consistent probability estimates from human subjects is very problematic: previous work in the literature on subjective error models and possible applications of quantized probabilistic labelling are discussed.

References

1. P. Smyth, 'Learning with stochastic supervision, ' in Computational Learning *Theory* and Natural *Learning* Systems 3, T. Petsche, M. Kearns, S. Hanson, R. Rivest (eds), Can bridge, MA: MIT Press, to appear, 1993.
2. P. Smyth, 'The nature of class labels for supervised learning,' in *Proceedings of the Fourth International Workshop on AI and Statistics*, Fort Lauderdale, Florida, January 1993.
3. B. Efron, 'The efficiency of logistic regression compared to normal discriminant analysis,' *Journal of* the *American Statistical* Association, vol.70, no.352, pp.892- 898, December 1975.

Topic: Lear-llillg Rules and Generalization